THE UTILITY OF PLEASURE IS A PAIN FOR DECISION THEORY*

EVERYBODY agrees that we have intrinsic desires. Obviously, we have desires; many of them are derivative; so, in a kind of analogy to a deductive theory that divides its provable statements into axioms and theorems, some of them must be basic.

There is much less agreement about what our intrinsic desires are or should be. Still, no one doubts that we intrinsically desire, among other things, happiness, pleasure, and the absence of pain, certain qualities of our sensations, feelings, moods, and psychological states in general. We desire these things, but without having, or feeling the need to have, a justification for doing so.

Desires come in various strengths, and, together with more or less firm beliefs, they guide rational action. The formal counterparts of these concepts in decision theory are utility and probability. The distinction between intrinsic and extrinsic desires is also reflected in decision theory, where each model contains a utility function that is basic (at least relative to that model) and an expected utility function that is derived from the basic utility function and the subjective probability function of the model.[1] On the whole, decision theory certainly provides the best formal explication of our intuitive reasoning about desires and beliefs in practical deliberation.

A person's intrinsically desired psychological states should thus be in the domain of her utility function. But then, we claim, decision theory fails. More specifically, we shall defend two theses:

(a) If a decision situation exhibits a certain causal structure, then

* We gratefully acknowledge the opportunities to discuss this paper at the Freie Universität/Berlin, the University of Tilburg, and the University of California/Irvine. In particular, we are very much indebted to Jack Birner, David Gauthier, Bert Hamminga, Dan Hausman, Isaac Levi, Gregory Kavka, Karel Lambert, Alan Nelson, Rainer Trapp, and Ernst Tugendhat. This work has been partially supported by the Deutsche Forschungsgemeinschaft, Grant No. Sp 279/2-1.

[1] In the classical theory of L. J. Savage, *The Foundations of Statistics* (New York: Wiley, 1954, 2nd ed. 1972), consequences have basic or absolute utilities, and actions get expected utilities. In the standard framework of decision trees [e.g., H. Raiffa, *Decision Analysis* (Reading, MA: Addison-Wesley, 1968), ch. 2] one may assume a basic utility either for each node or only for each path of the tree. By analyzing the tree by backward induction, one then assigns to each node an expected utility with respect to the subtree starting at that node. In the somewhat different theory of R. C. Jeffrey, *The Logic of Decision* (Chicago: University Press, 1965, 2nd ed. 1983), finally, basic utilities or nonprobabilistic values are assigned only to complete consistent novels or possible worlds, and propositions consisting of more than one possible world have expected utilities.

0022-362X/92/8901/10–29

decision theory is in trouble, because the derivation of expected utilities fails.

(b) This causal structure in fact obtains in a specific, but very common kind of situation, namely, when the intrinsically evaluated psychological states are in the domain of the utility function.

Section I explains thesis (a), section II defends thesis (b). Since the trouble we have concocted may seem to rest somehow on a misunderstanding, section III addresses this suspicion. We do not have a solution to the problem presented below and, indeed, we are not at all sure in what such a solution would consist. Nevertheless, in the final section some constructive conclusions will be adduced.

We shall talk about practical deliberation and decision theory only in an informal or semiformal way, because we shall discuss a general problem that is not produced by any one of the existing formal versions of decision theory. Our considerations are most easily formalized in terms of decision trees or decision-flow diagrams, however, as they are presented, in an exemplary way, in Raiffa (*op. cit.*). Thus, this theory should be used as a formal guideline, if required, and even though our informal intentions are broader, our claims about formal decision theory refer only to it.

It will be apparent that the problem is but a variant of Joseph Butler's[2] criticism of hedonism. Thus, in a sense, the point of our paper is that modern theorizing about practical deliberation has not dealt seriously with Butler's criticism.

I. A CAUSAL ASSUMPTION IN DECISION THEORY

Thesis (a) claims that decision theory presupposes a hidden assumption about the causal structure of decision situations.[3] To uncover it, let us look at how we proceed in practical deliberation.

The task is to find an action that optimally promotes our aims. For this purpose, we list our aims, i.e., the propositions we intrinsically desire to be true, and the acts open to us; each aim has a utility and, for each act, a probability of turning out true conditional on that act. Since each act thereby gets an expected utility, we know which acts have maximal expected utility, i.e., optimally promote our aims.

The calculation of the expected utilities of acts seems to proceed

[2] *Fifteen Sermons Preached at the Rolls Chapel*, 1729, in W. R. Matthews, ed., *Butler's Sermons and Dissertation on Virtue* (London: Bell, 1949), Preface §§ 29–31, and Sermon XI, §§ 1–10. Cf. also H. Sidgwick, *The Methods of Ethics* (New York: Macmillan, 7th ed. 1962), who deals with this criticism to some extent in book I, ch, iv, and in book II, ch. iii, § 5.

[3] There is, in fact, at least one other assumption of this kind. Newcomb's problem has given rise to differing versions of decision theory which converge only when some special conditions are assumed to hold. Cf., e.g., the papers in R. Campbell and L. Sowden, eds., *Paradoxes of Rationality and Cooperation* (Vancouver: British Columbia UP, 1985).

in one step. But usually it does not; usually, deliberation refers to a richer causal picture of the world that draws on the various ways and means by which our acts may influence our aims. This is reflected in a slightly fuller description of deliberation.

The task is to determine a total utility for each act and indeed, as we shall immediately see, for each proposition in consideration. A very natural assumption—which cannot be adopted as a definition because, as will soon emerge, it is false—is that the total utility $U^t(A)$ of a proposition A is simply the sum of its intrinsic utility $U^i(A)$ and its extrinsic utility $U^e(A)$.[4] Note that a proposition may have both: for some people, money not only makes for a comfortable life, but also acquires an intrinsic magic; driving is fun, but one also runs a risk of getting killed by it.

So deliberation starts from the agent's intrinsic utilities of all the propositions in consideration (many of them will be zero, of course). Then, by a kind of recursive procedure, the extrinsic and total utilities of each proposition are determined. The extrinsic utility of a proposition A is an expected utility; it is the weighted sum of the total utilities of the possible direct causal consequences of A, the weights being the subjective probabilities of these consequences conditional on A. By rolling his causal picture backward from its end points, the agent can thus successively calculate an extrinsic utility and, with the assumption just mentioned, a total utility for each proposition, until he has reached the starting points of his causal picture, i.e., his acts. He thereby finally determines which acts have maximal total utility.

However this sketch of practical deliberation is precisely formalized,[5] its essential point is the conception of desires or utilities and the derivation of them it embodies. There are only intrinsic and extrinsic utilities and sums of them; the intrinsic ones are underived, the extrinsic ones derived, and the derivation proceeds backward from supposed effects and their conditional probabilities and total utilities to supposed causes and their extrinsic utilities.

[4] Of course, each utility function always is that of some person at some time, even though we do not make this explicit in the notation.

[5] The sketch well agrees with common versions of decision theory. The simple one-step procedure is embodied in Savage's initial representation of decision situations in terms of states of the worlds, acts, and consequences (ch. 2). The more detailed multistep procedure is formalized in decision trees and their analysis by backward induction (see, e.g., Raiffa, ch. 2) which is, by the way, already contained in Savage's more elaborate theory of small worlds (sect. 5.5); our considerations are aligned with this standard formalization. The causal character of practical deliberation is, of course, accounted for in all of causal decision theory which, admittedly, is an elaboration of Savage's theory; cf., e.g., Brian Skyrms, *Pragmatics and Empiricism* (New Haven: Yale, 1984), ch. 4.

Talking thus of derivation may suggest that we conceive of decision theory as a syntactic theory about how rationally to compute utilities (and probabilities). This is not so. Decision theory should rather be conceived as stating laws about how utilities and probabilities coexist in rational complexes of graded beliefs and desires. What we mean, then, by using the more graphic phrase that some part of such a complex is derivable from other parts is that the former is uniquely determined by the latter according to the decision-theoretic laws.

Now, if practical deliberation is to work as described, some assumptions of causal well-orderedness have to be satisfied. One natural assumption is that the agent's complex of beliefs and desires entering his practical deliberation is a sufficient cause of his actual behavior. This assumption is, per definition, true of all rational agents.[6] And it is presupposed by practical deliberation in the sense that the weaker the impact of the deliberation on actual behavior, the emptier the deliberation itself. The anomalies that arise with violations of this assumption are well-discussed under the heading "weakness of will."[7]

Another assumption, and the one with which we are here concerned, is that the output of each step of deliberation is not already needed as input of that or an earlier step; otherwise, deliberation would obviously be caught in a circle. The output of deliberation consists in extrinsic and total utilities. How could they possibly be needed as input? When and only when the having of these extrinsic and total utilities is causally relevant to those effects from the utilities of which these extrinsic and total utilities are to be derived.

Let us look at a simple abstract instantiation of this seemingly weird possibility. Suppose that the agent's total utilities of B and non-B are already determined and that A is the only proposition considered by her to be directly causally relevant to B and non-B. Then the derivation of the extrinsic and the total utility of A is straightforward:

(1) $U^e(A) = U^t(B)P(B|A) + U^t(\text{non-}B)P(\text{non-}B|A)$
(2) $U^t(A) = U^i(A) + U^e(A)$

Now grant, for the sake of argument, that in the agent's view B causally depends also on the total utility she actually assigns to A. In this case, the probability $P(B|A)$ in (1) will not do; the deliberation

[6] At least for the causalists in action theory who hold that practical reasons may cause actions and indeed do so for rational agents.

[7] Cf., e.g., Donald Davidson, *Essays on Actions and Events* (New York: Oxford, 1980), essay 2; David Pears, *Motivated Irrationality* (New York: Oxford, 1986); and Kusser, *Dimensionen der Kritik von Wünschen* (Frankfurt: Athenäum, 1989), sect. 3.1 and 4.2.

must rather use the probabilities $P(B|A$ and $U^t(A) = x)$,[8] for any possible value x of $U^t(A)$. But then the decision-theoretic account of practical deliberation obviously gets circular; in order to derive the value of $U^e(A)$ according to the modified (1), one should know from which value of $U^t(A)$ to proceed, which according to (2), however, depends on the value $U^e(A)$ takes. Hence, for this account to work we have to assume the absence of such a vicious causal dependence.

We will claim, of course, that this dependence is neither weird nor impossible. But first the argument just given requires four comments on the abstract level.

First, the argument implicitly contains the assumption that what is causally dependent in the agent's view is also stochastically dependent according to her subjective probabilities. We believe this assumption to be defensible; but here is not the place to argue this intricate point.[9]

Second, the argument rests on the claim that the probabilities $P(B|A$ and $U^t(A) = x)$ must be used because $P(B|A)$ is unavailable. But this seems false. Why not make $P(B|A)$ available by computing the sum (or the integral) of all $P(B|A$ and $U^t(A) = x) \times P(U^t(A) = x|A)$ for all x? The idea of having a probability distribution for one's own total utility of A (this is what $P(U^t(A) = x|A)$ amounts to) looks a bit strange, however. More importantly, the idea is self-defeating, because that distribution would allow one to calculate first $P(B|A)$ and then, by (1) and (2), $U^t(A)$ and thus to gain certainty about $U^t(A)$. Of course, this self-defeat just reflects the circularity of the situation imagined.

The only way to escape this self-defeat would be to start for sure from such a value of $U^t(A)$ that will be confirmed by these calculations. This is the solution we shall suggest and more fully explain in section IV below. But note that this cannot be called a way of deriving $U^t(A)$ because there need not be exactly one such self-confirming value of $U^t(A)$. And even if there is exactly one such value, the procedure results in a determination of $U^t(A)$ that is very different from the one given by (1) and (2) in the absence of the vicious causal dependence. Indeed, in order to grasp fully how drastically the standard decision-theoretic picture is changed by adopting that solution, we first need to understand how that circular situation might come about (see section II) and why one cannot get rid of it within that standard picture (see section III).

Third, there is another tacit assumption in (1) and (2), namely,

[8] Here, '$U^t(A) = x$' stands for the proposition that the agent's total utility of A is x.

[9] For our view on the relation between causality and probability, cf. Spohn, *Grundlagen der Entscheidungstheorie* (Kronberg/Ts.: Scriptor, 1978), sect. 3.3 and 5.1; and "Direct and Indirect Causes," *Topoi,* IX (1990): 125–45.

that the intrinsic utility of A is independent of all the intrinsic utilities entering into the extrinsic utility of A. But this assumption may be dropped. If we allow for utility dependencies, we have to suppose the total utility of A-and-B and of A-and-non-B to be already given (either because the total utilities of these propositions are identical to their intrinsic ones, or because some steps of backward induction have already been performed). Working now one step backward in the decision tree, we obtain the total utility of A alone according to the formula:

(1′) $U^t(A) = U^t(A \text{ and } B)P(B \mid A) + U^t(A \text{ and non-}B)P(\text{non-}B \mid A)$

But grant again that B causally depends also on the total utility the agent actually assigns to A. Then (1′) gets into the very same trouble as do (1) and (2), for the very same reasons. Therefore, we gain some perspicuity and lose nothing by sticking to the mentioned assumption of utility independence; our problem about deriving extrinsic and total utilities hides in the probabilities and not in possible utility dependencies.

Finally, we have to be a bit more careful in specifying which causal assumption is required for (1) and (2) to work properly without circularity; for one might argue that the vicious causal structure is not in the least surprising. Trivially, the utility function one has is causally relevant to the act one performs[10] and thus also to everything affected by this act. But in our abstract example, for instance, this fact alone would not force practical deliberation to use the probabilities $P(B \mid A \text{ and } U^t(A) = x)$ instead of $P(B \mid A)$. If the causal influence of the actual value of $U^t(A)$ on B is screened off by A,[11] then these probabilities are the same for all x, and we need consider only $P(B \mid A)$. Thus, deliberation gets circular in this example only if the influence of the value of $U^t(A)$ on B is not screened off by A.

The causal assumption decision theory presupposes is therefore this: if the values of the extrinsic and total utilities of the propositions $A_1, \ldots, A_m$ are at all causally relevant to the propositions $B_1, \ldots, B_n$ from the utilities of which the utilities of $A_1, \ldots, A_m$ are to be derived, then this influence is screened off by $A_1, \ldots, A_m$ themselves. In each case in which this assumption is violated, the decision-theoretic derivation of expected or, in our terms, total utilities is caught in a vicious circle.

This consideration also makes it clear why this assumption is usually satisfied and why it may well have gone unnoticed. Our be-

[10] This is guaranteed by the above-mentioned assumption about the efficacy of practical deliberation.

[11] This means that, for all x, B is probabilistically independent of the proposition $U^t(A) = x$ conditional on A as well as on non-A.

liefs and desires certainly have a tremendous influence on the external world, but this influence is exclusively mediated and thus screened off by our acts.[12] Hence, the derivation of expected or total utilities of acts works smoothly as long as the propositions from the utilities of which the derivation starts refer only to the external world; and this is certainly true of most applications actually made of decision theory.

II. A VIOLATION OF THIS ASSUMPTION

Thesis (b) claims that this strange kind of causal situation which jumbles practical deliberation has a very common instantiation. The previous paragraph shows where to look for it; we have to consider our own inner states, in which way they are desired by us and in which way they are caused.

There are certainly many inner states having extrinsic utility for us; a noticeable example is given by beliefs that often have only instrumental value, e.g., for better informed decisions. But, more importantly, we also find a wide variety of sensations, emotions, moods, mentalities, basic frames of mind, and other inner states having intrinsic utility for us;[13] their desirability is of great practical impact and is not derived from other desires in any recognizable way. These states are classified by the occasions that produce them, by the ways in which they are expressed, by their very rich inner phenomenology, of course, and even by their desirability (unpleasant feelings, e.g., have to be undesired). Intrinsically desired inner states will be called *satisfactive states.*

How are satisfactive states caused? This is often opaque, as we all well know. But some general statements seem unassailable. Naturally, our satisfactive states are in the permanent causal grip of the external world. And naturally, the influence of the external world on us also depends on our inner condition it meets. For example, one usually enjoys an invitation with friends, but the more so, the less is one's stress. The crucial question is: What are the relevant parts of that inner condition? And the crucial answer is: among other things, how much we desire the external situations. That is, the inner state a person gets into is produced by the external situation one experiences and by one's prior inner state that includes her prior desire for that external situation often as a relevant part.

It is not difficult to find examples. Two men are watching Super

[12] This is an exaggeration, of course; our beliefs and desires are expressed not only in intentional action, but also in unintentional behavior (this fact is most beneficial to our transparency). But as long as we consider ourselves only as agents, as decision theory does and as we here do, we may stick to this exaggeration.

[13] This is not to exclude that these states may have extrinsic utility besides; but nothing of what follows turns on this possibility.

Bowl XXII on television. The Broncos have just scored the first touchdown. One man exults, the other gets nervous. Why? Because the one wants the Broncos to win, the other the Redskins; thus a touchdown by the Broncos has positive utility for the one and negative utility for the other.

Two years ago, a woman got pregnant, and she was miserable. Now she is pregnant again and very happy. Why? Because the first time she did not want to have a baby, but now she wants it. This is a causal explanation. First, there was the desire to have a baby, then she got pregnant, and now she feels happy; but if the desire had been absent, the pregnancy might not have made her happy, as was the case two years ago.

Prior utility and posterior satisfaction need not parallel each other, however. All kinds of relations are possible. The importance of an aim may make me pursue it so grimly that in the end I cannot enjoy the fruits of my efforts. Your badly wanting some kind of thing you are deprived of may work either way: you may not get enough and end up in total frustration, or you may humbly enjoy small amounts of it. There are things or situations that satisfy one only when one does not try to bring them about, or even when one does not desire them.[14]

All these examples exhibit the causal structure that the previous section argued to be problematic:

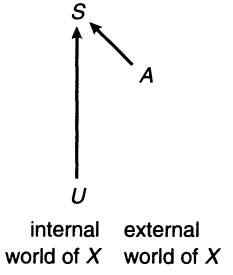

Figure 1

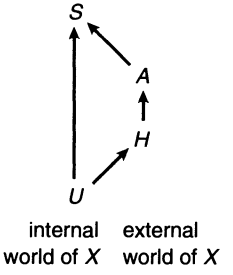

Figure 2

S: a satisfactive state of the person *X*

A: an external state of affairs

H: an act of *X*

U: *X*'s inner state of having a specific total utility function that has *S*, *A*, and *H* among its arguments

→ means: is causally relevant to

Figure 1 shows the bare essentials of this structure: they are realized in the football example (with *A* being the fact that the Broncos have just scored and *S* the exultation or, respectively, the depression). There is one causal chain running from the external fact *A* (via television, etc.) to the inner state *S*. And there is a second causal chain running from *U* to *S* which never leaves *X*'s internal world. Its

[14] Cf. Jon Elster, *Sour Grapes* (New York: Cambridge, 1983), sect. II.2, for a discussion of this kind of paradoxical situations.

details are not shown in the figure; but it is this chain which provides the psychological basis for how X takes up the outside happenings, and U contributes to this chain. There is no mention of any acts being contemplated in the example or in figure 1. But the essential point is already contained in this case, namely, that there is a second causal chain giving rise to the derivational circle unfolded in the previous section; hence, the extrinsic and thus the total utility of A cannot be derived from the intrinsic utility of S.

Figure 2 explicitly considers an act and fits the pregnancy example (with A being the woman's state of pregnancy, S her happiness, and H, e.g., the act of stopping the use of a contraceptive). Now, there are indeed two causal chains running from U to S, the first via H and A, and the second as before. As stated in the previous section, the first chain poses no problems because the influence of U mediated by this chain is screened off by H; therefore, the extrinsic utility of H, e.g., can be derived from the total utility of A. It is again the second chain not screened off by the external world which prevents the total utility of A, and thus that of H, from being derivable from that of S.

Intuitively, deliberation does not get off of the ground at all if the only desire it is based on is the desire to be happy. If you are trying to decide which profession to choose, for instance, you cannot arrive at a conclusion simply by trying to figure out how satisfying the various options would be—not because the decision would be so complex, but because the options as such do not have a definite satisfactoriness for you. Rather, you first need to have at least a rough idea of how much you want to belong to the various professions; only then can you reasonably ask how much they would satisfy you. This is essentially Butler's criticism of hedonism over 250 years ago:

> The very idea of an interested pursuit necessarily presupposes particular passions or appetites, since the very idea of interest or happiness consists in this, that an appetite or affection enjoys its object. It is not because we love ourselves that we find delight in such and such objects, but because we have particular affections towards them. Take away these affections, and you leave self-love absolutely nothing at all to employ itself about; no end or object for it to pursue, excepting only that of avoiding pain (*op. cit.*, Preface, §31).

We propose theses (a) and (b) as a translation of this into modern terms. If plausible, they point to an important incompleteness in decision theory. The gap does not open in each case in which our satisfactive states are part of the deliberation. Sometimes, the deliberational circle is side-stepped by, say, overriding moral considerations; sometimes the influence of the causal chain within the per-

son's internal world is negligible; and it would be most interesting to look at these exceptions more carefully. But in many ordinary cases the circle remains.

III. HOW NOT TO DENY THE CIRCLE

Although the present puzzle for practical deliberation may seem easily invented, there seems to be no easy way out. The following five suggestions, at least, will not do.

(A) It may be thought that we have relied too heavily on the intuitive causal picture of practical deliberation and its formal representation in decision trees, and that the circle vanishes in some other suitable formalization of decision theory. But this idea does not seem plausible if the candidate formalization is one of the current versions of so-called causal decision theory because they have the same roots as those to which we have referred. One might refer instead to some kind of so-called evidential decision theory; but we are skeptical whether this helps. In any case, the burden of proof does not seem to lie with us.

(B) One might think that we have simply offered an old refutation of hedonism in a new disguise, that only the most stubborn still believe in hedonism, and that we are beating a dead horse. This, however, would be a misunderstanding; our problem is a more general one. We do not merely criticize the hedonistic claim that the only things intrinsically evaluated are satisfactive states. Rather, our concern is with the usual decision-theoretic explanation of what it means to derive extrinsic from intrinsic values; and for that purpose we needed only to assume that some of the things intrinsically evaluated are inner states of ourselves. Thus, everybody who adheres to this explanation is in trouble, not just the hedonist.

Still, one might say that the problem emerged within a hedonistic venture, i.e., in the attempt to derive extrinsic utilities of external states of affairs from intrinsic utilities of satisfactive states. So why not simply abandon this attempt? But, surely, this is not yet an answer. The utilities of external states of affairs somehow relate to the intrinsic utilities of satisfactive states, even though this relation is not decision-theoretic derivation. We should not deny this relation, but try to account for it.

(C) Another suggestion is that the puzzle somehow is generated by an ambiguous use of the term 'utility'. In particular, one might think that, in the terms of figure 1, the utility of A that is causally relevant to S—let us call it $U^c(A)$, for the moment—is not the same as that utility of A that is to be derived from that of S. So what is $U^c(A)$? Is it really $U^t(A)$, as we have said?

One might suggest that $U^c(A)$ is a kind of evaluation different from the kind used in practical deliberation. But it need not be, of

course. We do not want to deny that propositions (or other things) are evaluated by us in other, e.g., aesthetic respects and that these other kinds of evaluation may be causally relevant to our satisfactive states. But we insist that the decision-theoretic kind of evaluation used in practical deliberation has this causal relevance, too; and all the examples given confirm that claim. (Even in the football example, the desires of the television watchers actually are of no practical relevance because they cannot do anything for their favorite teams; but they would, if they could.)

One might still object that there are now three alternatives for $U^c(A)$: $U^c(A)$ may be taken as the intrinsic, the extrinsic, or the total utility of A. $U^c(A)$ might indeed be $U^i(A)$; but then there is no deliberational circle because $U^i(A)$ is underived and available in advance. $U^c(A)$ cannot be $U^e(A)$; if it were and if we are right about the circle, there would not exist any derivable extrinsic utility of A. Finally, it seems that $U^c(A)$ cannot be $U^t(A)$ either; $U^t(A)$ was assumed to be the sum of $U^i(A)$ and $U^e(A)$, and if $U^e(A)$ does not exist, $U^t(A)$ does not either. Thus, the objection concludes, it seems to be an utter mystery which kind of utility should generate the circle.

There is a formal answer. In anticipation of this objection, we have emphasized in section I that it is only an assumption, not a definition, that $U^t(A)$ is the sum of $U^i(A)$ and $U^e(A)$. If this assumption is dropped—as is forced by the circle—the last step of the objection fails, and the identification of $U^c(A)$ with $U^t(A)$ can be preserved.

Our examples provide also an intuitive answer. The woman's happiness *is* influenced by how strongly she desires a baby; your satisfaction *is* influenced by how strongly you desire the profession you choose. In these cases, the desire affecting later satisfaction clearly is not (or not only) an intrinsic desire. In the woman's case, it is her total desire for the baby which springs from many sources: the intrinsic desire for the baby, the disutility of the various costs, the desire to please her husband or even her parents, etc., and, of course, also the desire for the satisfaction she derives from the baby. In the case of the profession, it is easily conceivable that the profession is in no way intrinsically desired, but only meant to satisfy, to yield earnings, etc.

Hence, the right conclusion from the objection is that there is something wrong not with our intuitive description of the circular cases, but with the usual classification of utilities into intrinsic, extrinsic, and total ones. The examples point to a kind of total utility which, though not intrinsic, cannot be decision theoretically derived in the presence of the circle. What we need is a more adequate classification of utilities.

(D) One might still feel, however, that the puzzle arises because it is not really clear what is meant by utility. The puzzle heavily relied on the causal wedge driven between utility and satisfaction; and this distinction may appear illegitimate. Indeed, many have tended somehow to equate the one with the other. We have to discuss separately the two ways of construing pronomial reference here; but it should be clear what we are up to: if you (1) reduce utility to satisfaction, then you do not do justice to utility; and if you (2) reduce satisfaction to utility, then you do not do justice to satisfaction. There is a way (3) of bridging the difference between utility and satisfaction with which we sympathize; but, as we will see, it does not resolve our deliberational circle either. Let us look at (1), (2), and (3) a bit more thoroughly.

(D-1) The agent, as depicted in decision theory, is quite a strange individual. He is always very busy calculating utilities and trying to maximize expected utility, and he seems to be done when having done so. But does he ever get anything? What is his ultimate pay-off? In the simple case where money takes the role of utility, the answer is easy: what he gets is money. In the real case where utility must not be equated with money, it is natural to give a similar answer: what he gets is utility. Indeed, what sense is there in trying to maximize the expectation of a quantity one does not get? So far, however, utility is only a name for the agent's ultimate pay-off. Now we know, of course, what his ultimate pay-off is: it is how he feels and how he is; in our terminology, it is his satisfactive states measured on some scale as conceived in the old calculus of pleasure. *This* is what utility is.

That is, briefly, the consideration that leads naturally to reducing utility to satisfaction; and many have been tempted by it.[15] But it does not yet give a complete account of utility. It only says that the agent's future utility pay-off consists in his future satisfactive states. But how are we to understand his present utility function? There seems to be only one way: the present utilities of his possible future satisfactive states are just the degrees of satisfaction they would give him were they realized; and the present utilities of other states of affairs represent expected satisfaction, i.e., they are expected utili-

[15] This reduction is characteristic of classical utilitarianism, but it may also be found in modern welfare theory. Cf. Jeremy Bentham, *An Introduction to the Principles of Morals and Legislation*, J. H. Burns and H. L. A. Hart, eds. (London: Athlone, 1970), ch. I; J. S. Mill, *Utilitarianism*, in *Collected Works*, F. E. L. Priestley, ed. (Toronto: University Press, 1969), vol X, pp. 203–59; Sidgwick, book II; and A. C. Pigou, *The Economics of Welfare* (New York: Macmillan, 1920), pt. I, ch. II. See also A. K. Sen, "Plural Utility," *Proceedings of the Aristotelian Society*, LXXXI (1980/81): 193–215; and J. Griffin, *Well-Being* (New York: Oxford, 1986), ch. 1.

ties representing the supposed conduciveness of these states to future satisfaction.

This conception is nevertheless inadequate. Its picture of practical deliberation and rational agency contains only the agent's possible future satisfactions and his present beliefs about them, while totally dispensing with the agent's desires or wants. We take it as obvious that such a picture is fundamentally distorted.

One may try to improve this picture by locating the agent's desires in it, i.e., by defining them within it. This, however, looks unpromising from the beginning, because desires are widely taken as a basic kind of propositional attitude not reducible to other propositional attitudes.[16] Moreover, the only way of locating them seems to be to say that the agent's present utilities, which have already been defined, simultaneously represent his present, more or less intense desires. But this move turns hedonism into an a priori truth; it stipulates in effect that the agent's desires are intrinsically concerned *only* with his future satisfactive states. Moreover, the need of fitting this stipulation to the intuitive notion of desire entails a traditional problem of hedonism, namely, that of having to relate each and every desire to the experience of satisfaction and each and every preference to a difference in (the expectation of) the experience of satisfaction. Again, all this seems distorted.

Thus, the crucial point is, in a nutshell, that present desires may not be identified with present expectations about future satisfactions and that present utilities must represent the former and not the latter in order to play the role for practical deliberation and rational agency which they actually play in decision theory.[17]

Finally, it must be noticed that this conception of utilities does not make the deliberational circle disappear; it only transforms it. Our claim (b) then amounts to the claim that the agent's present expectation of future satisfaction is causally relevant to future satisfaction; and it is supported by precisely the same examples. It says, in simpler nonprobabilistic terms, that the agent's present belief that this state of affairs will produce that future satisfaction is causally relevant to whether it actually is so. This kind of belief is, so to speak, self-verifying (or in some other way causally affects its own truth). And it generates an analogous kind of deliberational circle: to know which future satisfaction you should expect, you should know which future satisfaction you do expect. Hence, this conception of utilities is not only misguided, it is also unhelpful, and we should resist it.

[16] See also D. Lewis, "Desire as Belief," *Mind*, XCVII (1988): 323–32, for a general argument why desire is not reducible to belief.

[17] This is more thoroughly argued in Kusser, sect. 3.3.

(D-2) In fact, the history of the concept of utility in this century is shaped by such resistance, as is reflected, e.g., in the so-called theory of revealed preferences. Most thinkers, in particular economists, nowadays conceive utilities as representing various desires of varying strength as they are revealed by the preferences and actions of rational persons. But then the question of the relation between utilities and satisfaction re-emerges.

There has been a long-standing tendency to reverse the answer just discussed, i.e., to say that satisfaction is whatever we have when our desires come true and thus to turn satisfaction into a purely formal notion.[18] Probably no one has put this so bluntly because it sounds so absurd. But the tendency to conceive satisfaction (or pleasure or happiness) in this way has ample historical precedent. For example, you find Immanuel Kant[19] saying:

> Glückseligkeit ist der Zustand eines vernünftigen Wesens in der Welt, dem es im Ganzen seiner Existenz alles nach Wunsch und Willen geht (*ibid.*, p. 224).

And John Rawls[20] saying:

> . . . we are happy when our rational plans are going well, our more important aims being fulfilled, and we are with reason quite sure that our good fortune will continue (*ibid.*, p. 548).

Nevertheless, we should resist any such tendency. We have got what we wanted too often without being satisfied; and the phenomenology of what we have called satisfactive states is too vivid and too independent. So, no conceptual assimilation of satisfaction to realization of desires can be successful.

There are, of course, reasons for this tendency. One reason certainly is that it is at least the normal case that the realization of our desires satisfies us; but not every case is normal. A stronger reason is that in each case we have to assume that the realization of our desires will satisfy us. If we believe that some of our desires will not satisfy us and if we still stick to these desires, then we either have overriding, say, moral reasons for doing so, or we are plain foolish. The situation resembles G. E. Moore's paradox of belief: we have to believe that what we believe is true; but others may know better, and we may learn to know better, too. Likewise, we have to believe that what we desire will satisfy us. But others may know better, and we may learn to know better, too. Such reasons may have furthered this tendency, but they do not justify it.

[18] This is nicely paralleled by a purely formal notion of pay-off utility as something you get or have whenever a desire of yours gets fulfilled. Cf. Griffin, ch. 1, and Sen, sect. 4.

[19] *Kritik der praktischen Vernunft*, 1788.

[20] *A Theory of Justice* (Cambridge: Harvard, 1971).

(D-3) Once we acknowledge that the relation between satisfactions and desires or utilities, i.e., the grounds for rational action, is contingent, the deliberational circle stands. There is some support for the view that, in some sense, our genuine or *true* desires are those which actually satisfy us and that we ideally have such desires.[21] We share this view, and we will expand on it below. But it has no impact on the deliberational circle. This circle is generated by the relation between actual desires or utilities and satisfactions and does not vanish by considering how desires ideally are.

(E) You might still feel uneasy about the theses leading to the circle. Perhaps you are tempted by the following objection. "Crudely put, your thesis (b) says that very often a person enjoys something because she wants it, and then you somehow extract a problem from this. But if you ask this person why she wants that thing, it would be natural for her to reply: 'Because I enjoy it.' And there is nothing wrong with that reply—except that you cannot accept both the reply and your thesis (b); you cannot have it both ways."

One can easily be misled by this consideration; indeed we have been. There is, however, no incoherency at all; the two 'because's in that objection simply have two different meanings. When we say that a person enjoys something because she wants it, we state one of the many causes of her enjoyment. When she says that she wants it because she enjoys it, she expresses one of the many practical reasons for her want.

It is almost as simple as that, but not quite; two problems hide in this answer:

(E-1) We have just admitted that the person's belief that she will enjoy something is a practical reason for her desiring it. But what is a practical reason? Is A not a practical reason for B just in case A is essentially used in a practical deliberation supporting B, i.e., in a decision-theoretic derivation leading to B? The problem is that the last sentence contains in fact two different explanations; the first is vague and correct, the second precise and incorrect. The conclusion

[21] Cf. Bertrand Russell, *The Analysis of Mind* (London: Allen & Unwin, 1921) ch. III, esp. pp. 72ff.; J. C. B. Gosling, *Pleasure and Desire: The Case for Hedonism Reviewed* (New York: Oxford, 1969), ch. 7; Ernst Tugendhat, *Probleme der Ethik* (Stuttgart: Reclam, 1984), pp. 33–56, esp. pp. 45f.; and A. Tversky and D. Kahneman, "The Framing of Decisions and the Psychology of Choice," in Elster, ed., *Rational Choice* (New York: University Press, 1986), p. 138.

It would not be correct to require of a rational person to have true desires; we also do not require of her to have true beliefs. Seeking, not having knowledge is essential to rationality (of course, the second often is the result of the first). Thus, it certainly belongs to rationality to strive after true desires. This sense of rationality may be intended in the above quotations of Kant and Rawls.

we have to face is that there are two kinds of practical reasons: the decision-theoretic kind and some other kind. This other kind is instantiated by the belief that one will enjoy something, but it still lacks an explanation.

(E-2) The causalists in action theory, with whom we have obviously sided, hold that, for a rational person at least, what is a practical reason from the person's inside perspective turns into a cause from our outside perspective. Now, when we admit that the person's belief that she will enjoy something is a practical reason for her for desiring it (and think that she is rational), we also have to say that her belief causes her desire. And since we keep claiming that her desire is causally relevant to her actual enjoyment (and think that causal relevance is transitive), we have to conclude that her belief that she will enjoy something is causally relevant to its truth. Should we really be prepared to accept this conclusion?

IV. HOW TO TAKE THE CIRCLE SERIOUSLY

It seems that the circle stands unshaken, but each attempt to avoid it opens questions. If the circle is important, it presumably calls for a theory. We have none; but we want to suggest some constructive observations and some positive conclusions.

We have hinted at the rudiments of a formal solution of our problem already in section I; let us spell them out by looking again at formulas (1) and (2). In these equations, the relation between A and B was assumed to be regular, so to speak. In the circular case, however, which can only obtain when B represents a future satisfactive state, the causal beliefs of the agent are expressed not by $P(B \mid A)$, but rather by $P(B \mid A$ and $U^t(A) = x)$, for any possible value x of $U^t(A)$. Therefore, the former should be replaced in (1) by the latter. Now, if $U^t(A)$ were known to take the value x, the modified (1) and (2) would determine a value y of $U^t(A)$. Thus, if the value of $U^t(A)$ is x, it should be y; and this is satisfiable, only if $x = y$. In other words, only solutions of the equation

$$(3)\quad x = U^t(A) + U^t(B)P(B \mid A \text{ and } U^t(A) = x) + U^t(\text{non-}B)P(\text{non-}B \mid A \text{ and } U^t(A) = x).$$

are eligible as values of $U^t(A)$; otherwise, the total utility of A does not match the expected satisfaction A derives from B, given this total utility of A.

The fact that the unknown x appears both at the left hand and the right hand side of (3) nicely reflects our deliberational circle. Note that no extrinsic utility is now assigned to A; when total utility is constrained by (3), it can no longer be assumed to be the sum of intrinsic and extrinsic utility. Equation (3) may have no, exactly one, or, as will be quite common, more than one solution, depending on

how the conditional probabilities in (3) vary with x. If the mathematical problem were neatly formalized, one could presumably bring fixed-point theorems of equilibrium theory into action. We have not attempted to do so, for reasons stated below.

Nevertheless, the equilibrium idea embodied in (3) is intuitively very helpful; it informally clarifies six of the questions we have left open.

Four kinds of utility. Equation (3) specifies the relation between the utilities of satisfactive states and those of external states of affairs which was sought in (B) of the previous section. This relation complements the decision-theoretic picture of practical rationality. According to this picture, there are only two kinds of utility: utilities figuring as derivational premises, i.e., the intrinsic ones, and derived utilities, i.e., the extrinsic and total ones. In (3), however, utilities have different roles. It would be inappropriate to say that in (3) $U^t(A)$ is derived from $U^t(B)$ or from $U^i(B)$.[22] This is particularly obvious in the case where (3) has more than one solution. Thus, the intrinsic utilities of satisfactive states like B do not in general function as derivational premises. Correspondingly, the utilities of many external states of affairs like A are neither derived nor intrinsic in the original sense. Rather, the relation between $U^t(B)$, or $U^i(B)$, and $U^t(A)$ in (3) is more appropriately described as one of *control*. The utilities of external states of affairs are controlled by those of satisfactive states in the sense that the former have to be in equilibrium with the latter and have to be changed whenever they are out of equilibrium.[23] Utilities may also be subject to such an equilibrium control—this is the essential lesson to be drawn from our deliberational circle.

In this sense, (3) generates two new kinds of utility: controlling and controlled utilities. This also clears up the muddle in (C) about which kind of utility of A is causally relevant to satisfaction. It is the underived total utility of A—where 'underived' does not mean "intrinsic" in the old sense, and 'total' does not mean "derived." There is nothing paradoxical in this assertion, once the four kinds of utility have been recognized.

Two kinds of practical reasons. What we have just said also answers the question in (E-1) of the previous section. There are indeed two kinds of practical reasons: those functioning as premises in a

[22] $U^t(B)$ and $U^i(B)$ will usually be equal if B is a satisfactive state, as it is in (3).

[23] Indeed, the relation is less one-sided than the word 'control' suggests. One may re-establish equilibrium also by re-evaluating satisfactive states. Thus, when you experience that the fulfillment of a certain desire does not satisfy you, you may also try to learn to draw satisfaction from it, e.g., by redirecting your awareness, by depreciating alternative satisfactions, etc. Indeed, this method of accommodating pleasures to desires (instead of desires to pleasures) may often be observed. It should not be dismissed as irrational.

decision-theoretic derivation and those supporting equilibria like solutions of (3). The desire for satisfaction and the belief that one will enjoy something—or, more generally, utilities and probabilities like the ones on the right hand side of (3)—are reasons of this second kind. Thus, if asked why one wants something, there is nothing wrong with the answer: "Because I enjoy it." What would in general be wrong is the assumption that this want is simply derived from the desire to feel joy.

The causalists' claim that practical reasons viewed from inside are causes viewed from outside has to be specified accordingly. It seems clear that this claim can only be maintained for the decision-theoretic kind of practical reasons for which it was intended. The practical reasons of the equilibristic kind at best cause one to stick to the desire supported by them, but not to have it in the first place. We thereby can avoid the undesired conclusion of (E-2); but the next point looks into the matter more carefully.

The status of the belief that the realization of a desire will satisfy. What we have just said means that this belief accompanies rather than precedes the desire. Indeed, it is a necessary companion of the desire. Unless one has other reasons for a desire, one would be irrational to have the desire and lack the belief that its realization will satisfy. This or rather the probabilistic counterpart thereof is what (3) asserts; and this agrees well with what we have observed in (D-2).

What about the suspicion lurking in (E-2) and also at the end of (D-1) that this belief is not a normal empirical belief, but rather one of the self-verifying kind? We have to be careful here concerning the precise content of this belief. It may be (i): the realization of a desired proposition, say, A, will satisfy. Or it may be (ii): the realization of the proposition A as a desired one will satisfy. Anyone who believes (i) will probably also believe (ii); and anyone who believes (ii) and is conscious of his desire for A will certainly believe (i). Still there is a subtle difference. If one's causal belief is expressed by (i) and one overlooks or denies that one's desire for A is causally relevant to how satisfying A is, then one can maintain the received picture—at the cost that the belief in (i) is in fact causally affecting its truth. If, however, one's causal belief is expressed by (ii), one has to give up the received picture with respect to the utility of A—with the benefit that the belief in (ii) is a normal empirical belief. Only this second alternative can be sustained. The first is self-defeating; as soon as one realizes that one's belief in (i) is affecting its own truth, one can no longer stick to it.

Changes of the belief that the realization of a desire will satisfy. When we said that this belief necessarily accompanies the desire, this did not mean that this belief would merely be a by-product of the

desire. On the contrary, under reading (ii) it is a normal belief to which normal epistemology applies. It may be grounded or ungrounded by evidence; it may be false; one may learn that it is false; and, to put it so generally as to cover also the probabilistic case, one may change it on the ground of new evidence according to general rules of rational belief change. But which consequence has such a change for the desires? Again, we have to distinguish two cases.

The normal case is that belief change only calls for a recalculation of the expected utilities. This may even be true of changes of the beliefs about future satisfaction. For instance, I might decide in a restaurant to try an exotic dish that I have never tasted before. It turns out that I do not especially like it, and I accordingly revise my belief about the satisfying efficacy of that kind of dish and thus its expected utility. This case allows such a simple description, however, only when our circle is absent, i.e., when one's desire for a dish does not affect how much one likes its taste.

The case is very different when our deliberational circle is present. The change of the beliefs about future satisfaction, i.e., of the probabilities involved in (3), then results in a predicament: the controlling desire for satisfaction and the controlled desires for other things are out of equilibrium and can no longer be maintained rationally. Somehow, they must be changed; somewhere, a new equilibrium must be found; but there is no hint about how to change them and where to seek it. Everyone will know the experience of running on empty—e.g., when the thrill of rock 'n' roll fades or when one's love is worn out—and the often difficult process of redirecting one's desires.

Equation (3) as a model of deliberation. Equations (1) and (2) provided rules of practical deliberation; they told us how to derive extrinsic and maximize total utilities. Equation (3), which replaces (1) and (2) in the cases discussed, might thus also be taken as offering a model of deliberation, namely, this one: solve (3); if it has no solution, that is too bad (though this may be ruled out by suitable continuity assumptions); if it has a solution, fix $U^t(A)$ to be the maximal solution—because among all the eligible values of $U^t(A)$ this is the one which gives you the highest expected satisfaction! If this were a feasible model of deliberation, then it should be helpful to work out the mathematics. But we doubt this very much, for two reasons.

First, the relevant probabilities seem scarcely to be available; for knowing them means having an answer to all the questions of the following kind: "If I should desire A so and so strongly, how likely is it that the realization of A will satisfy me to such and such a degree?" But we simply do not know and thus cannot grade many of the

possible satisfactions. How does it feel to be a free climber? How satisfying would it be? We have not the slightest idea. And it is even more difficult to take into account the influence of the desire. How satisfying would it be to be a free climber, if one really wanted to be one? Thus, we can plausibly answer these questions only for the kinds of satisfactions we have experienced and for the kinds of desires we have pursued; all other probabilities would be artificial guesswork.

Secondly, the formation of aims and ends, i.e., of underived desires is quite obscure; the idea that one could adopt, by sheer will, so to speak, the desires that such deliberation recommends, looks a bit strange. Thus, it seems doubtful whether such deliberation could really be as effective as it should be in order to have a point.

Rational and true desires. If (3) should not be understood as a model of deliberation, what does it then accomplish? It provides a model of *learning*—at least in the weak sense that it says that desires in disequilibrium should rationally change and that this change should rationally move to some new equilibrium.

Indeed, (3) establishes an extended notion of rational desire. In the received picture, there was no way at all of assessing intrinsic desires or utilities as rational or irrational; this had a precise meaning only for extrinsic desires. In our amended picture, however, (3) provides a standard of rationality also for controlled as well as for controlling desires (and thereby opens a possibility of grasping the notion of rational feelings). This is tantamount to saying that (3) provides extended means of criticizing the desires of a person, not according to moral or otherwise external standards, but strictly by that person's own measures.

Insofar as the beliefs that the realization of desires will satisfy may be true, or insofar as the probabilities in (3) may be the objective ones, (3) finally gives some meaning to the notion of *true* desires: our true desires are those which are in a maximal equilibrium given true beliefs. We rationally search for true desires; and since the causation of our satisfactions keeps changing, we have to search again and again. This is what we in fact do. We do not gain practical help from (3), but perhaps some theoretical insight into what we are seeking.

ANNA KUSSER

Freie Universität/Berlin

WOLFGANG SPOHN

Universität Bielefeld